MUTTINENI, SHRAVYA, M.S. Estimating Sentiment in Social Media - a Case Study of the Migrant Caravans News on Twitter. (2021)
Directed by Dr. Jing Deng. 38 pp.

Social media has empowered us to gain immense knowledge, criticism, and positivity from every corner of the world. Accessing social network sites like Twitter and Facebook has slowly become the norm, as people are now inclined towards social media for news and opinions rather than reading traditional newspapers or watching TV news. There are many instances when a piece of news in these sites, irrespective of its credibility and reliability, has a significant impact on people's opinions towards it. Social Media has created a platform for many individual users, small business owners and large co-operations to make a living by targeting their users and generating leads through digital Marketing strategies. Although these networks being extremely useful in many ways from the recent events we have seen a greater influence of these platforms in one's life effecting mental well being and creating a social dilemma.

In this work, we focus on identifying such problems of social media posting sentiment efficiently with the use of estimation and references through the methods of sentiment analysis for tweets. We calculated the sentiments of all the tweets related to the migrant caravan issue. Using the scores from these tweets, I have analyzed the outcomes.We used Vader Sentiment Analysis to analyze the sentiment of each tweet before estimating the sentiment on Twitter toward this news. We used different weights for the tweets in particular weeks. Estimated the effects it could have on the users from the computational analysis we performed. we came through different kinds of visualizations to present out work. we even tried some machine learning algorithms targeting better efficiency and automated results but, left most of these for our future work. We investigated some queried tweets of around 40 weeks of a sensational tragic issue of "Central American migrant caravans" and how the

tweets played their critical roles in the overall sentiment on the twittersphere. We show that such simple estimates can be very accurate.

Our study can be helpful in its efficient estimation of social media content sentiment analysis. This includes potential identification of user cluster and inherent communities, dynamic community detection on-the-fly, community structure migrations, etc.

ESTIMATING SENTIMENT IN SOCIAL MEDIA - A CASE STUDY OF THE MIGRANT CARAVANS NEWS ON TWITTER

by

Shravya Muttineni

A Thesis Submitted to
the Faculty of The Graduate School at
The University of North Carolina at Greensboro
in Partial Fulfillment
of the Requirements for the Degree
Master of Science

Greensboro
2021

Approved by

Committee Chair

To Shashank,

For his advice, patience and strong faith in me.

To my Dad,

Muttineni Satish Kumar

For his continuous support, for having blind trust in me and for always respecting every decision I made.

To my Mom,

Tirumani Sreedevi

Thank you for always being there for me!

APPROVAL PAGE

This thesis written by Shravya Muttineni has been approved by the following committee of the Faculty of The Graduate School at The University of North Carolina at Greensboro.

Committee Chair _____
Jing Deng

Committee Members _____
Lixin Fu

_____
Minjeong Kim

_____
Date of Acceptance by Committee

_____
Date of Final Oral Examination

ACKNOWLEDGMENTS

I want to thank my guide, Jing Deng, from the bottom of my heart for assisting me with planning this Thesis. His experience and prompt were priceless; this has allowed me to present a conference paper for SCSN conference 2021. The freedom he has given me to explore the various aspects always gave me strength to make decisions and do what I am best at. I want to thank Dr.Deng once again for all that he has done for me.

PREFACE

As the world moves further into the digital age, the use of social media has become an integral part of our daily life from accessing news and information, socialising to decision making. While social media continues to endlessly impact our abilities to convey, build relationships, access information; From the recent events we also find a greater part of these platforms contributing towards conspiracies, manipulation and influence in politics, addiction and not to forget its toll on one's mental health. The significant part of this research focuses on investigating and examining what part of Social Networks is actually healthy and what adverse effects does it possess on one's daily life. I overviewed for conclusion examination to help my exploration on informal communities and the impact it has on individuals in the general public. The impact of social networks has a huge role to play in our life. There have been so many instances where the world economy also depended on the trends of social media. Social media also has bought in social dilemma in people around the world. The emotional aspects including mental decisions and conditions are also playing a prime character. I was curious on the off chance that I could work on the system and accomplish comparable or better outcomes utilizing few strategies.

TABLE OF CONTENTS

LIST OF TABLES

Page

LIST OF FIGURES

Page

CHAPTER I

INTRODUCTION

I.1. Overview

Social networking media play immensely vital roles in our daily lives. Most of the time, we get carried away by the news circulated here. It is crucial to not just believe in the current affairs published there. It is also equally essential to check if all the odds are right while being considerate about the situations. However, with the increasing usage of social media, it has become very tough to contain any news and restrict people from being influenced by the aura that it carries. Most of the time, getting to the root level of the news which is spreading rapidly is also tricky. There are many reported cases of death because of wrong blame or the negativity spread by social media, especially in celebrities [1]. There are a few essential considerations while getting carried by social media news: Are all users tweets about an issue the same? Should people disregard all the negative tweets? Which one or ones should be trusted more? Are all the positive tweets real? What is the pivotal tweet that is influencing other people the most?

These were a few of the questions that motivated us to investigate user tweets, and we focused on Twitter data due to several reasons. Twitter is a free social networking micro-blogging service that allows registered members to broadcast short posts called tweets. Twitter members can broadcast tweets and follow other user's tweets by using multiple platforms and devices. Twitter usage positively influenced knowledge acquisition most of the time [1].

" Privacy is not a privilege; it is a fundamental right " but current trends of social media although having set privacy policies in place these site collect enough personal data of

their user thorough cookies etc to track their activity levels to an extent that if an AI model is created from this data it can easily predict the users real life responses / reactions to any scenarios. Although certain sites enforce strict policies to restrict access of private data we have had enough data breaches lately to say that Privacy has become a novel concept. Mere Social media presence or existence of any user means no more privacy. In the recent events some SNS's like Facebook have enforced even more stricter guidelines to protect user data and avoid from future breaches.

[] Offline Social Network is simply a hardware device which acts as a rendez-vous point between various users located in the area of reach of that device, who can potentially form a social network, exchange data, store their own data, use the local data stored, while potentially enjoying all the standard functionalities of Online Social Network, such as chat, walls, etc. The fact that there is no connection to the Internet and that the accesses to that box are anonymous by default (depending on the users and the parameters), the users privacy is preserved (provided the box is trusted). The users may interact through the Offline Social Network and create an ephemeral social network, reflecting the fact that they are at the same location at the same time while not implying any longstanding digital relationships in the future. The information stored on the Offline Social Network can be temporary as a guarantee that the data durability is limited over time and that no user can be traced. It can also be permanent for instance when associated to a specific location.

I.2. Importance of Social Networks

Social networks initially were used to build strong relationships, bonding between people, communicate, work for each other and also would help contribute lot to the society. But, the new era has defined Social networks as virtual communication sites that allow its participants to connect, build relationships, and collaborate on social issues. It became part of our lives and spread rapidly among youth. Young people join these sites to keep strong

relationships with friends and to make new ones. Therefore, it is important to investigate the factors that influence the intention to use social networking sites (SNSs) to gain better position in the social reform among young people [2]

In the course of recent years, we have seen the rise of new worldview in the Internet, online web-based media organizing, which allow Internet clients to convey and work together with family, companions, gatherings of people, and other local area by utilizing online media instruments (i.e., Twitter, Facebook, Instagram, Snap-chat, Tik-Tok, Youtube, Reddit, Pinterest etc). The utilization of social media for correspondence is getting more predominant around the world, with individuals from nations of fluctuating financial advancement progressively getting to the Internet to partake in systems administration locales.

Online web-based media organizing today is an extraordinary platform to meet and coordinate with individuals sharing comparative business interests, platforms like Meetup provide an opportunity to engage and build professional and personal relations among people with similar interests. Be that as it may, they can likewise present genuine security dangers to clients and their associations.

The use of social networking is getting more pervasive around the world, with individuals from nations of changing monetary improvement progressively getting to the Internet to take part in systems administration destinations. With the prevalence of portable gadgets and applications joined with interpersonal interaction innovations, correspondence utilizing on the web long range interpersonal communication instruments is turning into another lifestyle to individuals.

Along with the most common communication and social relation and opinion building sites like Twitter, Facebook, Linkedin there several other types of social networking ranging from image sharing sites, Content Curation , Content blogging, Discussion Forums,

Community blogs.

The most widely used type of Social Media majorly in current younger generations mainly the Millennial's is Media sharing Platforms. Media sharing type is a social platform whose sole purpose is to engages its users through sharing media content like photographs, videos, live video, and other kinds of media on the web. These types are currently being used by most small business organizations to large cooperations for digital marketing in brand building, generating leads, by targeting audience and so on. Most popularly used sites for this type are Instagram, SnapChat, Youtube. The most popular users of these platforms are termed as influencer's and with the help of these people most companies market their products to a target audience and widen the customer base.

Discussion Forums are another commonly used type of social media channels basically used for finding and sharing information, asking question and discussing or having debates and share an individuals experiences and opinions on any topic or genre. These discussion forums have a massive number of users and it ensures unprecedented reach. These kind of experience sharing can be quiet helpful for other people going through similar circumstances, they can gain knowledge or learn from other persons experiences and take some precautionary steps if necessary. These forums provide the answers to different queries of any domain. Although these are quiet helpful in many scenarios they are some circumstances that can off lead to a non-healthy note. The recent events of hate crimes, heated debates, criticism and negativity have been allover the internet. Most widely used platform for this category are apps or site like Reddit, Quora etc. Way before Facebook and Twitter existed these were the platforms widely used by professionals, experts and enthusiasts for having debates and can also be very eventful in advertising and running digital marketing campaigns.

Review Based Networking sites are the ones that will help you find out, share your

views and opinions on different information about a variety of products, services or brands. These review based sites can be beneficial to businesses having positive reviews on these networks, their claims turn more credible because reviews on these networks act as Social Proof. In the current market trends it is very important for any organisation to have a positive review on these sites and also by resolving all the disputes raised by your customers on these platforms is equally important. Yelp, Trip Advisor, Trust Pilot etc are some examples of such types of social media platforms.

Then there are Content Curation Platforms like pinterest. These types of Social Media will help you find variety of latest content and media that are trending. They are very helpful in channelizing and curating you projects by creating mood boards or planning boards that you can use to visualize the outcome of your plans before you actually execute them.

Content Blogging is a type of Social network for publishing, discovering and commenting on articles, blogs and other content on the web. Content marketing is one of the most powerful ways to target, attract, engage and convert a target audience. WordPress and Blogger are the traditional blogging platforms while Tumblr (micro-blogging service) and Medium (Social Publishing Platform) is the latest blogging and publishing networks. These networks are must for the associations that need to effectively use Content Marketing, besides, you can share this content on an arrangement of Social Networks like Facebook, Twitter, LinkedIn etc, Content that you use on these networks will moreover help you with making a strength for businesses and groups who are searching for information stressed that claim to fame will indeed visit your blog or website page.

I.3. Twitter Sentiment Research

Twitter is one of the most grounded development interpersonal interaction administrations and right now in excess of 500 million clients appreciate it for sharing data,

expressing strong opinions and delivering information on current affairs in seconds.

The impact of twitter social networking site has a huge influence on the people especially on the millennial's. knowingly on unknowingly it has started impacting right from the roots. Information spreads like a wild fire, campaigns are strongly held, opinions are expressed unconditionally and affects of this are reflects on almost all the industries. There are no barriers for expression but, how our views our impacting daily lives of people (celebrities to common people is something that needs to be considered) Twitter has made insurgency as well as it has likewise affected clients of all age gatherings, it has achieved elevated requirements among the other systems administration locales. The patterns in twitter has affected the world from the official decisions to viral recordings. Monetary and securities exchange declarations from twitter additionally impacts the stock costs. Twitter has been an effectively available open vehicle for big names, influencers to contact individuals better. Numerous missions, support for a reason, offer thoughts out boisterously. The majority of the occasions everyday person has a medium to communicate and has the chance to be heard from utilizing organizing destinations. As, everything has low and high, there are equivalent detriments as benefits.

CHAPTER II

RELATED WORK

II.1. Twitter Social Network

II.1.1. Analysis from Facebook and twitter data

In [3], He et al. used data from the pizza industry by applying text mining to analyze unstructured text content on Facebook and Twitter sites of the three largest pizza chains. They revealed the value of social media competitive analysis and the power of text mining as an effective technique to extract business value from the vast amount of available social media data. Recommendations were also provided to help companies develop their social media competitive analysis strategy.

In [4], Elbagir and Yang examined the sentiment analysis of Twitter data. They described how sentiment analysis concerned with public data like twitter can obtain and automatically classify their sentiment polarity. This was particularly interesting because of Twitter's shorter message/post style. In their study, Valence Aware Dictionary for sEntiment Reasoner (VADER) was used to classify the sentiments expressed in Twitter data. Contrary to the fact that many other previous studies were more inclined towards binary classification, VADER has shown good accuracy in detecting ternary, and multiple classes.

II.1.2. Sentiment Analysis using Emoticons and Hashtags

In [5], Hutto and Gilbert used a parsimonious rule-based model for sentiment analysis, where they constructed a gold-standard list of lexical features specifically tuned to sentiment in microblog-like contexts. They combined these lexical features with consideration for five general rules that embody grammatical and syntactical conventions for expressing and emphasizing sentiment intensity.

In [6], Wang et al. manually labelled tweet sentiments before assigning three raters to label the sentiments. Two of these raters were normal and one was senior. In conflicting situations, they would treat it with the senior rater. These information were then fed with retweet numbers of the tweets to construct a repost cascade tree, repost diffusion network, and sentiment reveal. Sentiment patterns would then be analyzed.

II.1.3. Impacts of social media on people

In [7], Allcott and Hunt explained about some of the welfare affects of social media. With their reseacrh and a randomized experiment they explaines how the rise of social media has provoked both optimism about potential societal benefits and concern about harms such as addiction, depression, and political polarization. In the experiment, by deactivating Facebook for the four weeks before the 2018 US midterm election and increasing other offline and active activities could actually cause a large persistent reduction in anxiety,stress and gave users some balanced and happy mental health.

In [8], Noor and Norhisham conducted extensive research in their article Systematic Review of Social Media Acceptance From the Perspective of Educational and Information Systems Theories and Models explained that the uses and gratifications theory and the social constructivism theory were considered the most widely used educational theories in social media. How each of these models could contribute to analysis of the educational and information systems theories that are used to examine the social media

In [9], Chandler and Ali collected data on 176 million tweets from 2011 to 2014 with content related to depression or suicide using the ARIMA data analysis model. Using this data they have computed the trends and analyzed the impact of social media. From their research they could conclude that Spikes in tweet volume following a behavioral health event often last for less than 2 days. Individuals and organizations that want to disseminate behavioral health messages on Twitter in response to heightened periods of interest need to

take this limited time frame into account.Also, their analysis model model offered an empirically based measure to identify periods of greater interest for timing the dissemination of credible information related to mental health.

In [10], Burnap and Omer gathered and broke down the information gushing from twitter,they have examined the chance of estimating spikes in friendly strain. Distinctive computational techniques were tested to recognize spikes in strain utilizing a human coded test of information gathered from Twitter, identifying with an allegation of racial maltreatment during a Premier League football match. Discussion examination joined with syntactic and dictionary based content mining rules; notion investigation; and AI techniques was tried as a potential methodology. The outcomes showed a mix of discussion investigation strategies and text mining beats various AI draws near and an opinion examination apparatus at grouping pressure levels in singular tweets.

II.1.4. Sentiment Analysis using ML models to handle Twitter Data

In [11], Wang et al. performed a sentiment analysis on the 2012 presidential tweets using a Naive Bayes algorithm. To create their training dataset, they employed AMT to label tweets for them. They used four categories (positive, negative, neutral, and unsure) and achieved 59% accuracy.

In [12], Wang et al. tried to automatically label tweets based on emotion. They used hashtags labels such as "#happy" and "#sorrow" to train a classier to identify tweets that expressed joy or sadness. They achieved an accuracy as high as 65.65%. In [13], Davidov et al. performed a variety of sentiment analyses using tweets. This included using hashtags as labels to train a classier to identify "focused" sentiment. For example, a hashtag that includes an emotion and a target such as "#tmobilesucks" could be used to calculate the sentiment that users express toward T-Mobile.

In [14], Gautam and Yadav extracted features from the dataset after basic

pre-processing then used these extracted features and classified using support-vector machines and Naïve Bayes.

In [15], Amolik et al. performed a sentiment analysis of movie reviews using machine learning algorithms such as Naive Bayes and support-vector Machine. They used 600 positive, 600 neutral, and 600 negative tweets in their experiment. They also replaced usernames with a generic marker "AT_USER."

In [16], Kolchyna et al. combined lexicon and machine learning techniques by using a lexicon scoring scheme as the input for the machine learning algorithm. A manually labeled set of tweets was used to construct a lexicon. This was accomplished by determine the number of times a word appeared in a positively or negatively labeled sentence. Each word was given a positive and negative sentiment score between 0 and 1. This input was used for experiments with Naive Bayes and Support-Vector Machine Algorithms.

In [17], Sentiment analysis of twitter, Apoorv and Boyi used a unigram model to two different types of classifications, positive versus negative and a 3-way positive versus negative versus neutral. They explained and visualized comprehensive set of experiments for both tasks on manually annotated data that is a random sample of stream of tweets. They investigated two kinds of models: tree kernel and feature based models to demonstrate that both these models outperform the unigram baseline. They have also explained feature-based approach in which feature analysis reveals that the most important features are those that combine the prior polarity of words and their parts-of-speech tags.

In [18], Mishra and Rajnish explained how opinions in tweets matter. They have performed opinion mining on various tweets considering the topic of elections in India.They collected the tweets using the twitter API and used dictionary based approach to analyze the data posted by the users. Then they classified the tweets based on the polarity of the tweet from the user. This has actually helped us to start with our research.

In [19], Anjaria and Guddeti investigated the sentiment prediction task over Twitter using machine-learning techniques, with the consideration of Twitter-specific social network structure such as retweet. They also concentrated on finding both direct and extended terms related to the event and thereby understanding its effect. They employed supervised machine-learning techniques such as support-vector machines (SVM), Naive Bayes, maximum entropy and artificial neural networks to classify the Twitter data using unigram, bigram and unigram + bigram (hybrid) feature extraction model for the case study of US Presidential Elections 2012 and Karnataka State Assembly Elections (India) 2013. Also, with their approach they could achieve upto 88 percent accuracy

II.1.5. Previous Research vs. Our Research

Compared to these works, we approach the tweet sentiment issue from a group perspective. We design an estimate mechanism for tweet sentiment that only requires the sentiment and retweet numbers of a very small portion of tweets related to the topic. This would allow our estimation to be rather accurate

The goal of this research is to develop a method to estimate the overall sentiment without the need to check the sentiments of all tweets or even the tweets themselves. Our case study focuses on finding out [2] the positive or negative sentiment of user tweets about the "Central American migrant caravans" [3]. Our approach is to use a VADER sentiment analysis. VADER (Valence Aware Dictionary and Sentiment Reasoner) is a lexicon and rule-based sentiment analysis tool specifically attuned to sentiments expressed in social media and works well on texts from other domains. We use Data Pre processing techniques from genesim to statistically remove redundant words from tweets, thus improving the quality of the tweets, run time, and accuracy. A similar approach can betaken with text sentiment analysis using AllenNLP [4].Our simple estimates are based on those tweets that are most influential on the topic. These results are then combined into a single sentiment

number that is between1and1, where1represents most positive and1most negative. In order to check estimate accuracy, these estimates are then compared to the sentiment measurement that is computed from all tweets bearing the same topic.

CHAPTER III

TWITTER SOCIAL NETWORK

III.1. Central American migrant caravans

Central American migrant caravans, also known as the Viacrucis del Migrante ("Migrant's Way of the Cross") are migrant caravans that travel from Central America to the Mexico–United States border. The largest and best known of these were organized by Pueblo Sin Fronteras (Village Without Borders) that set off during Holy Week in early 2017 and 2018 from the Northern Triangle of Central America (NTCA), but such caravans of migrants began arriving several years earlier, and other unrelated caravans continued to arrive into late 2018.

There is some disagreement as to whether the migrant caravans are primarily composed of refugees seeking asylum or are merely large concentrations of traditional economic migrants. Numerous human rights organizations document the increase in violence and abuse in recent years in Central American countries.A report by the Geneva Declaration on Armed Violence and Development, cited by Amnesty International, noted that between 2007 and 2012, several Central American countries had the highest average annual female homicide rates in the world, although the average annual male homicide rates in the world are higher. Other studies of the composition of the caravans indicated that the caravans more resemble traditional economic migrants. The causes of the migration, as well as the proper way to settle or deport the migrants themselves, remains a source of political debate within the U.S.

Twitter has reacted very strongly on this burning issue back then. People with around mixed emotions and feelings expressed their strong views on the happenings. There

have been few tweets which have effected the opinions about others and have a turn around reaction on this. This has almost continued for around 38-40 weeks in the twittersphere with weeks 20-25 to be the most active and highly burning.

III.2. Methodology

III.2.1. VADER sentiment analysis

We used a VADER sentiment analysis. VADER (Valence Aware Dictionary and Sentiment Reasoner) is a lexicon and rule-based sentiment analysis tool specifically attuned to sentiments expressed in social media and works well on texts from other domains. Gilbert [6] developed VADER, which is a simple rule based model for general sentiment analysis and compared its effectiveness to 11 typical state-of-the-practice benchmarks, including Affective Norms for English Words(ANEW), Linguistic Inquiry and Word Count (LIWC), the General Inquirer, Senti WordNet, and machine learning-oriented techniques that rely on the Naive Bayes, Maximum Entropy,and support-vector Machine (SVM) algorithms.we used a combination of qualitative and quantitative methods to produce and validate a sentiment lexicon that is used in the social media domain.VADER is utilizing a parsimonious rule-based model to assess the sentiment of tweets. the VADER lexicon performs exceptionally well in the social media domain. The correlation coefficient shows that VADER (r = 0.881) performs as well as individual human raters (r = 0.888) at matching ground truth (aggregated group mean from 20 human raters for sentiment intensity of each tweet) the VADER sentiment lexicon is gold-standard quality and has been validated by humans. VADER distinguishes itself from LIWC in that it is more sensitive to sentiment expressions in social media contexts while also generalizing more favorably to other domains. VADER (Valence Aware Dictionary and sEntiment Rea-soner) is a lexicon and rule-based sentiment analysis tool that is specifically attuned to sentiments expressed in social media [19]. VADER uses a combination of a sentiment lexicon that is a list of lexical features (e.g.,

words), which are generally labeled according to their semantic orientation as either positive or negative. VADER has been found to be quite successful when dealing with social media texts, New York Times editorials, movie reviews, and product reviews [20].This is because VADER not only tells about the Positivity and Negativity score but also tells us about its positive or negative sentiment. The Compound score is a metric that calculates the sum of all the lexicon ratings which have been normalized between -1(most extreme negative) and +1(most extreme positive). Between these two, there are positive sentiment (compound score >= 0.05), neutral sentiment (-0.05< compound score < 0.05), and negative sentiment (compound score <= -0.05).

III.2.2. Data Preparation

In this work, we used Tweepy an open source Python package that gives a convenient way to access the Twitter API with Python. Tweepy includes a set of classes and methods that represent Twitter's models and API endpoints, and it transparently handles various implementation details such as: aata encoding and decoding. Twitter's API allows us to do complex queries like pulling every tweet about a specific topic within the last twenty minutes or obtaining an individual user's non-retweeted tweets [20]. There is a limit on API calls. For a standard user it is 900 requests every 15 minutes. The Twitter dataset, which was mined from particular dates with keywords given (in our case study, those related to the migrant caravans.) Unlike other social platforms, almost every user's tweets are completely public and retrievable [21].

We first tried to analyze the data for our experiments and decide on those variables which are essential and related to our work. There are 2,793,838 tweets in our dataset, only a small number of which were removed due to bad formatting during the importing process. Using the genesim python library, we removed any graphical characters, multiple white spaces, punctuations, stopwords, numeric, and converted the text encoding to UTF-8.

Removal of stop words such as ("and," "or," etc.) was performed to get a cleaner version of the text. We left spaces alone and changed the text to lower for better accuracy and quality of the text. We used tweets of 40 weeks, which is close to around nine months of tweets related to the issue. The figure below represents the total number of tweets across the 40 weeks showing the oscillation of tweets between the weeks. Week 20-25 have been prime weeks where most of the users posed opinion on the issue.Later, it gradually decreased after a certain point. It typically shows a bell graph and start and ends being subtle.

III.2.3. Sentiment Calculation

We have first calculated the sentiment for each each tweet and labelled the negative sentiment to be -1 positive to be 1 and 0 for neutral. Below is a graph showing the positive and sentiments of users for a week across all the weeks. Just an overview on what the sentiments of the tweets are and the progress of reactions with the seriousness of the issue. Figure III.1 is a visualization of the entire tweet data for every week, showing both negative and positive tweets.

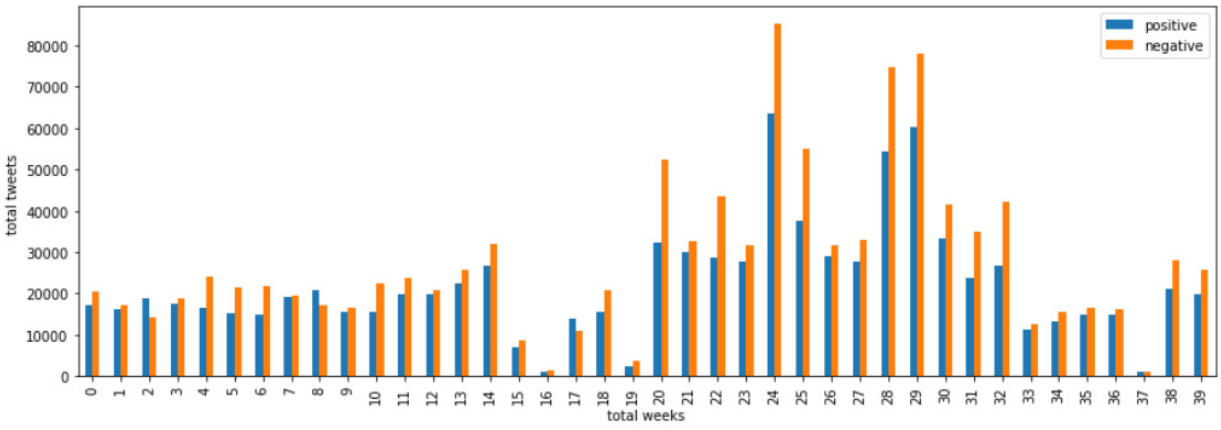

Figure III.1. Tweets with sentiments across all weeks

III.2.4. Influential Tweet

A retweet is a re-posting of a tweet. Twitter's retweet feature allows users to share tweets with followers quickly. In the retrieved data, the retweet column indicates the number of times a tweet has been retweeted. This would allow us to deduce the most influential tweets. There are other indications as well, but we leave them for our future work.

III.2.5. Computational analysis of sentiment's ratios

Computational analysis of ratios of positive and negatives is performed in this section to understand how the change in the rate contributes to summarizing the overall sentiment of that week.

For the k-th week, we compute the positive and negative ratio as follows:

$$r^+(k) = \frac{N^+(k)}{N^+(k) + N^-(k) + N^0(k)} \tag{III.1}$$

$$r^-(k) = \frac{N^-(k)}{N^+(k) + N^-(k) + N^0(k)} \tag{III.2}$$

where $N^+(k), N^-(k), N^0(k)$ are the numbers of positive, negative, and neutral tweets, respectively.

In addition, all tweets are not equal in their influences in the Twitter sphere. While there are other ways to measure the influence of a tweet, one straightforward method is to adjust a weight that is based on its retweets and favorites (similar to "likes" on Facebook). Thus, the new definition becomes

$$r^{+'}(k) = \frac{N^{+'}(k)}{N^{+'}(k) + N^{-'}(k) + N^{0'}(k)} \tag{III.3}$$

$$r^{-'}(k) = \frac{N^{-'}(k)}{N^{+'}(k) + N^{-'}(k) + N^{0'}(k)} \tag{III.4}$$

where

$$N^{\tau'}(k) = \sum_{i=1}^{N^{\tau}(k)} (1 + R_i + \alpha F_i) \tag{III.5}$$

where $\tau \in \{+, -, 0\}$, R_i and F_i are the number of retweets and favorites of tweet i, respectively, α is a weight parameter to be adjusted between retweets and favorites and we have counted the tweet itself as a retweet (hence the "1" term inside the summation terms. We will check the impact of different α values in Section III.3. We further searched and downloaded the timeline of about 20 news related to the migrant caravans and correlated these with our tweet sentiment analysis. These major news served as the source of major "jolts" to the Twittersphere sentiment dynamics in the studied period. Each of the 20 news has been analyzed as positive or negative toward the overall sentiment and each of these is then applied to the sentiment dynamics in the same week where the news broke in the more traditional news channel or the Internet.

We need a mechanism to represent the sentiment of entire group of users that shows a certain ratio of tweets with positive, negative, and neutral sentiments $(0 \leq r^{+'}(k), r^{-'}(k), r^{0'}(k) \leq 1)$. It is natural to represent the overall sentiment by subtracting the negative tweet ratio from the positive tweet ratio. We have two variables $(r^{+'}(k) - r^{-'}(k),\ r^{0'}(k))$ left. Because all three variables sum up to 1, the possible zone where the duo values can take can be shown in a triangle. It can be observed that with a large $r^{0'}(k)$, the possible range for $r^{+'}(k) - r^{-'}(k)$ is smaller. In general, the bottom right corner of the triangle is most positive and bottom left corner is most negative.

Therefore, we define "direct sentiment" and "weighted sentiment", the second of which we simply call "sentiment" as

$$\Psi(k) = r^{+}(k) - r^{-}(k) \tag{III.6}$$

$$\Psi'(k) = r^{+'}(k) - r^{-'}(k) \tag{III.7}$$

The definition in Eq. (III.7) ensures that

$$-1 \leq \Psi(k), \Psi'(k) \leq 1 \ . \tag{III.8}$$

With this definition, we are interested in finding out a small ratio of tweets in each block time that can be used to analyze and estimate the overall sentiment rather accurately.

Define a ratio call $0 \leq \xi \leq 1$. Suppose we collect ξ of the most influential tweets in the block of time and use these to compute the sentiment and we want to identify the value of ξ such that the estimation is not too far away from the sentiment that is computed from all of the tweets.

III.3. Experiments and Results

III.3.1. Data Preparation

Here is the cleaned version of the tweets to perform sentiment analysis.

One example of a raw data:

"people crying migrant children separated le migrant estÂ lâavenir duÂ families bug t care people s right freedom movement perfectly content monde http bit ly owtzl pic families sent away odniqcnay kept completely cages oppose freedom"

The corresponding pre-processed data:

"people crying migrant children separated families care people right freedom movement perfectly content families sent away kept completely cages oppose freedom"

Figure III.2 is a visualization of the entire tweet data for every week. A few weeks lack sufficient data because of crawl errors (week 4 and 40). Actual news appeared starting week 5-6.

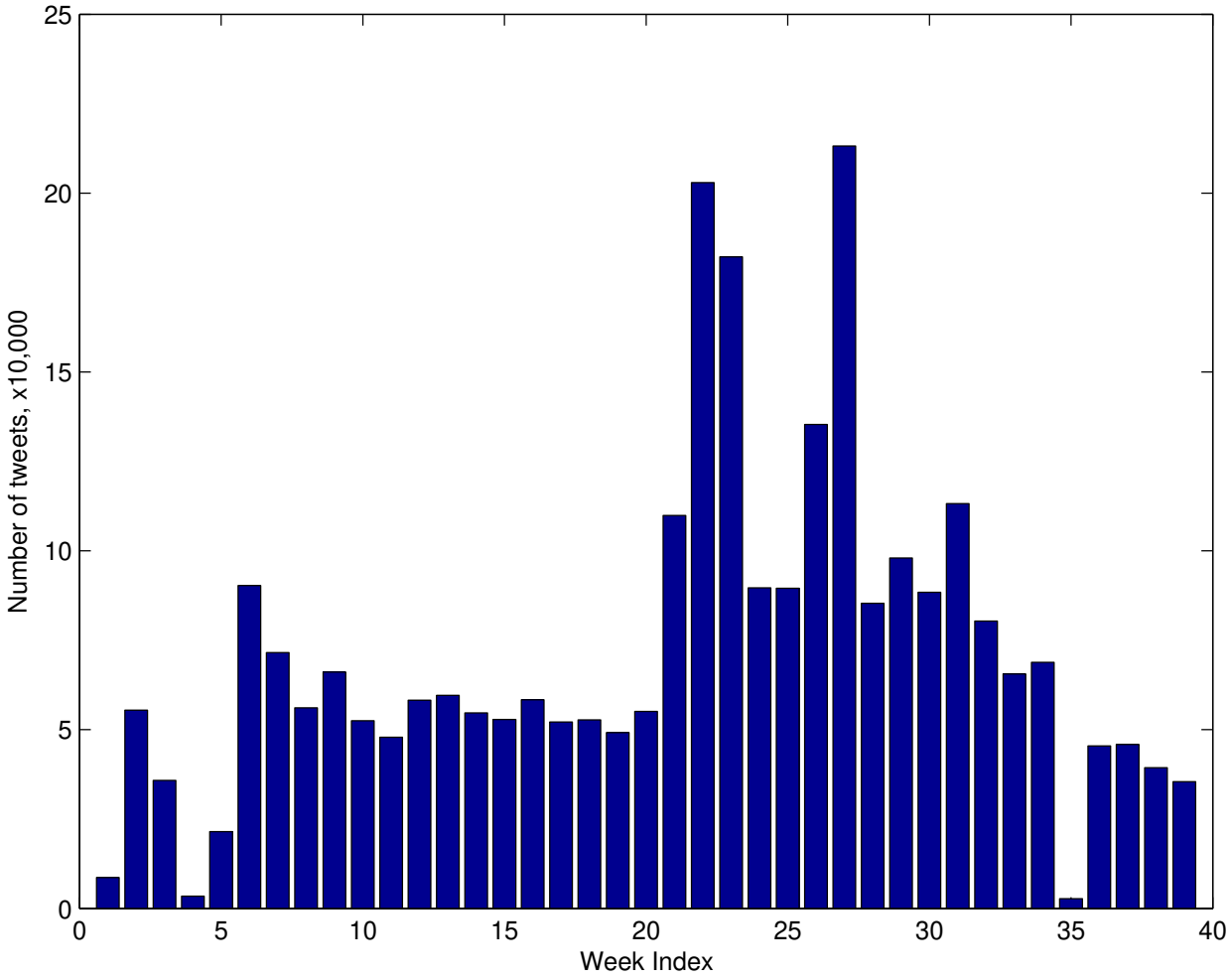

Figure III.2. Number of tweets for all the weeks

III.3.2. Influential tweets

These are a few tweets that have been retweeted a lot, seemingly impacting other users' mindset for the event:

Tweet #1: "tomi lahren said watching migrants hit tear gas highlight thanksgiving weekend people evil deranged psychopathic imagine getting suffering particularly suffering innocent children." This tweet has been retweeted 66,600 times and the sentiment of the tweet is negative.

Tweet #2: "these patriarchal okoumou black woman protested migrant family separations sitting foot statue liberty said come children released." This tweet has been retweeted 36,479 times and the sentiment of the tweet is positive.

An example of a rather neutral tweet is: "new migrant children remain separated parents month court s deadline handful reunited weeks deported parents reunited children." This tweet has been retweeted 3,077 times and the sentiment of the tweet is neutral.

Figure III.3 is the visualization results of positive, negative, and neutral sentiments with respect to the tweet count for each of the 39 weeks.

Furthermore, we randomly picked one of the accounts and ran some analysis on it. The user had a total of 503 tweets in 39 weeks. Most of the tweets were negative, and there was not a clear trend with progression in the weeks. Among all the tweets, 68% of them were negative sentiment, 25% of the tweets were positive sentiment, and a few were neutral.

In TABLE III.1, we showed details of a few more influential users and their tweets with its retweets and favorites count listed along with their sentiments.

III.3.3. positive and negative ratios for tweets

We have estimated the avg positive and negative ratios for tweets for a week. The average ratios varied from 0.46 to 0.03 with negative ratios taking the upper curve.

Figure III.4 is a visualization of the positive and negative ratios across 40 weeks.

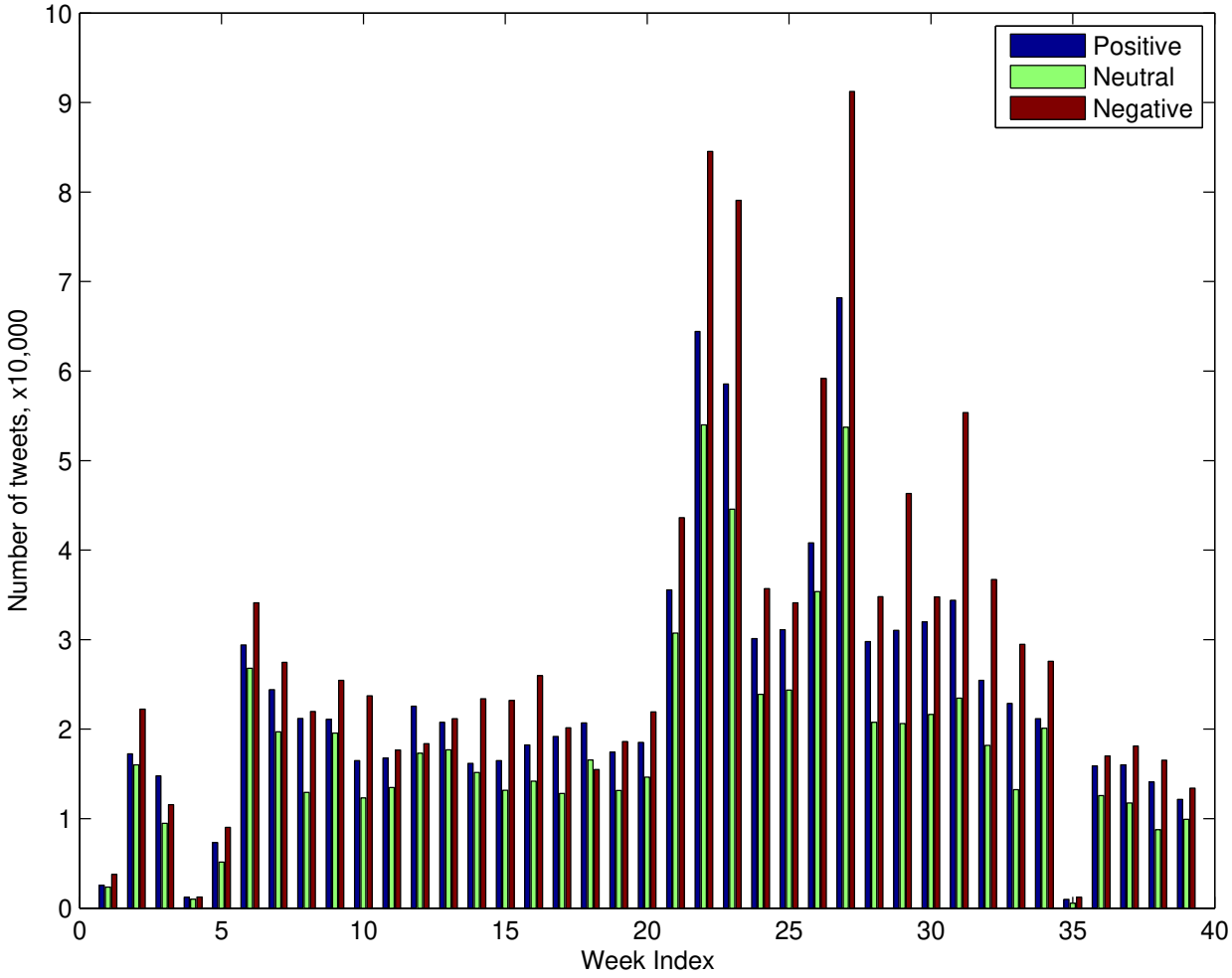

Figure III.3. Tweets with sentiments

III.3.4. positive and negative ratios for varying weights

Estimated the negative and positive ratios for tweets across all the weeks with varying alpha values, used different alpha values to score the ratios of the sentiments for the tweets.

Figure III.5 is a visualization of the positive and negative ratios across 40 weeks.

III.3.5. User tweet Sentiment

In Fig. III.6, we show the sentiment results of tweets by a single user, which was chosen randomly among some of the accounts with more tweets. A large fluctuation can be

Table III.1. Analysis of tweets with Retweets and Favorites along with the sentiment for 10 more influential users

Username	Retweet	Favorites	Sentiment
Dav*son	12539	29519	positive
The*one	9735	35011	negative
Am*ek	13132	20013	negative
Red*otr	31597	112888	negative
Jei*gn	36479	107523	positive
MrF*tik	66600	224208	negative
Din*uza	9399	30806	negative
Pas*ino	28749	83260	negative
Nyt*mes	20146	13834	negative
Rep*edy	15714	37349	negative

observed in the number of tweets posted by this user, so are the numbers of different sentiments. In some weeks, more than 25 negative tweets were posted, while no tweets on the topic were posted at all in other weeks.

III.3.6. Retweets sentiment estimates

Retweets are those tweets which are mostly followed by people on the twitter. A tweet is retweeted when the users have an opinion to spread the tweet to reach out the larger crowd. We have considered the most retweeted tweet from each week, and calculated the sentiment of the the retweet. The retweet sentiment value varied from -1 to 1. We can also conclude that the retweets sentiment might have a huge impact on the tweets for the week. There are evidences that this most retweeted tweets sentiment has created an impact to form an opinion on the issue.

III.3.7. Sentiment Estimate

As mentioned earlier, all tweets are not equal in their influence. we calculated the influence of the tweet by adjusting the weight based on favorites and retweets for a tweet.

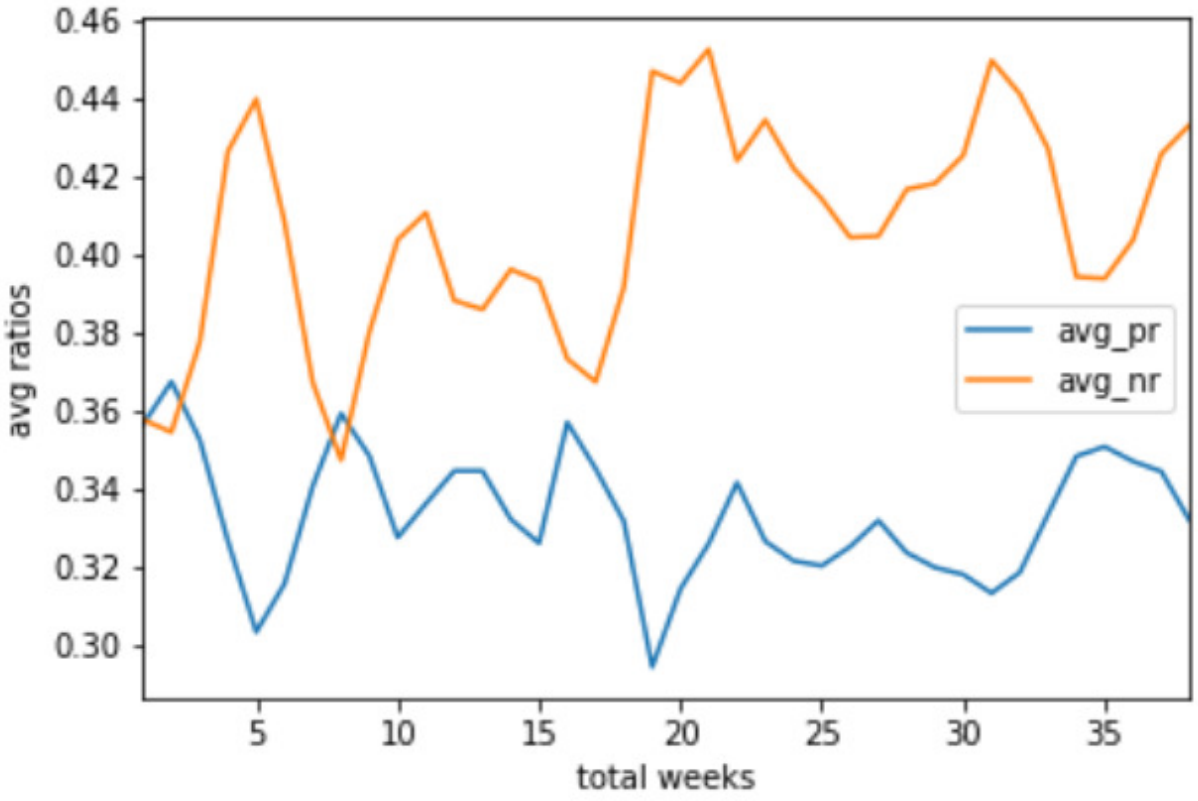

Figure III.4. Average ratios of sentiments

Here, we have used 3 different values for alpha to compare the best results among them.

Figure III.7 shows the results of positive and negative ratios with weighted parameters and varying α. This graph explains the fluctuations in the values and the difference between the ratios. Even when α changes from 0.1 to 10, sentiment results do not change much. We will use $\alpha = 1.0$ henceforth.

We show the results of different estimates based on different ξ ratios in Fig. III.8. Obviously, when $\xi = 1.0$, the estimate considers all of the tweets and the sentiment value is close to the real value. As ξ decreases from 1.0, accuracy of the estimates drops, as can be seen by some of the gaps in Fig. III.8.

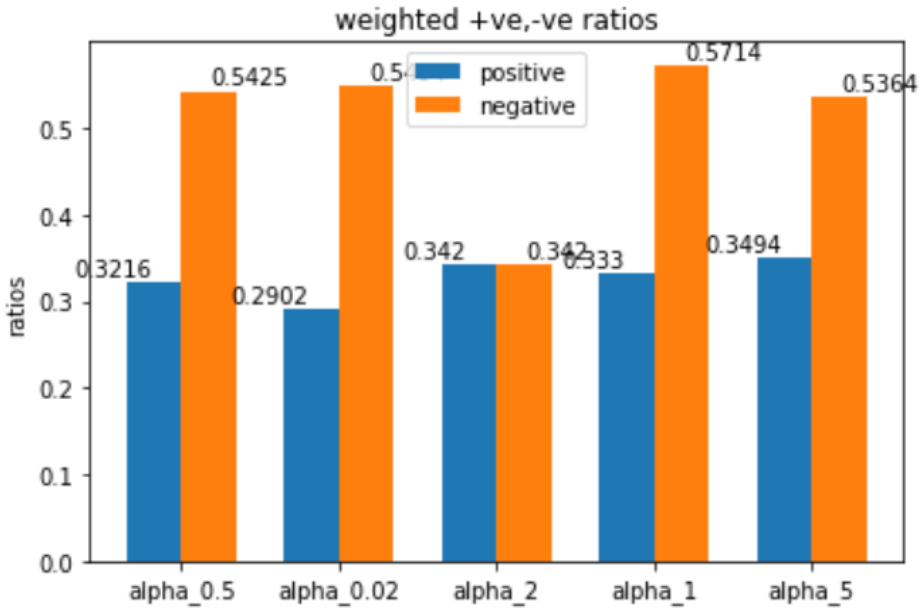

Figure III.5. Sentiment ratios with varying alpha values

In Fig. III.9, we show the estimate error as a function of ξ. Even with a relatively small ξ (such as $\xi = 0.001$), our estimate is rather accurate (estimate error is about 0.035 when $\xi = 0.001$ and it drops to 0.01 when $\xi = 0.01$). This is to say that even when we are checking the overall sentiment based on 0.1% of the all relevant tweets, our estimation of the sentiment is only about 0.035 away from the actual value.

The estimate error of direct sentiment Ψ based on ξ of tweets is also shown in Fig. III.9. While these errors are slightly higher than that of weighted sentiment Ψ', they are still very small, underlying the accuracy of our approach.

III.3.8. Sentiment Estimate for Retweets

We first mined the tweets that have been retweeted from the entire data. We matched the records of tweets and the accounts which retweeted this tweet. Once, we have the data regarding the same we located these tweets from the 2 million data set. Now, we

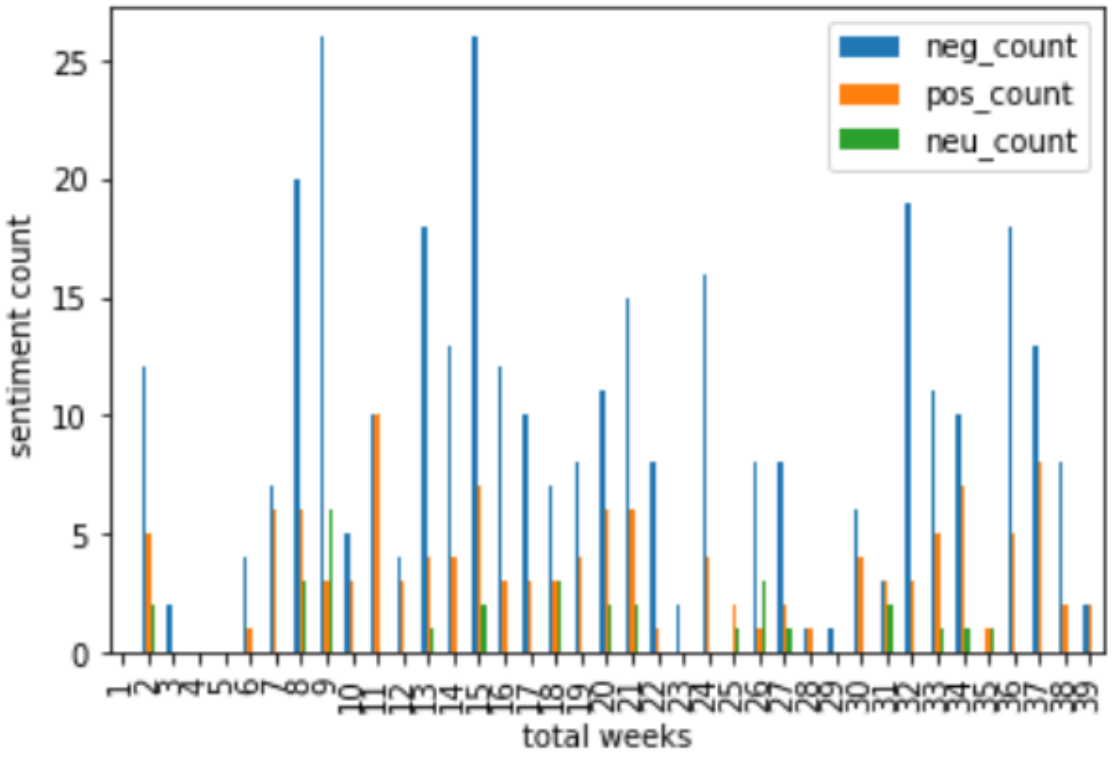

Figure III.6. Single user tweet sentiments for 40 weeks

have calculated how the averages of sentiments have changed before and after this retweet. There were many instances where a retweet has effected negatively in the tweets of other users. There are drastic changes in the averages of the sentiment values before and after a particular retweet has made. Another, important factor is we always had to consider was, there are few instances where the most liked tweet and retweeted tweet aren't the same. So, from this we could conclude that though a lot of people supported a particular tweet by liking they dint actually follow it.

Figure III.7 shows the results of average sentiment ratios before and after the most retweeted tweet. This graph explains the fluctuations in the values and the difference between the average ratios.

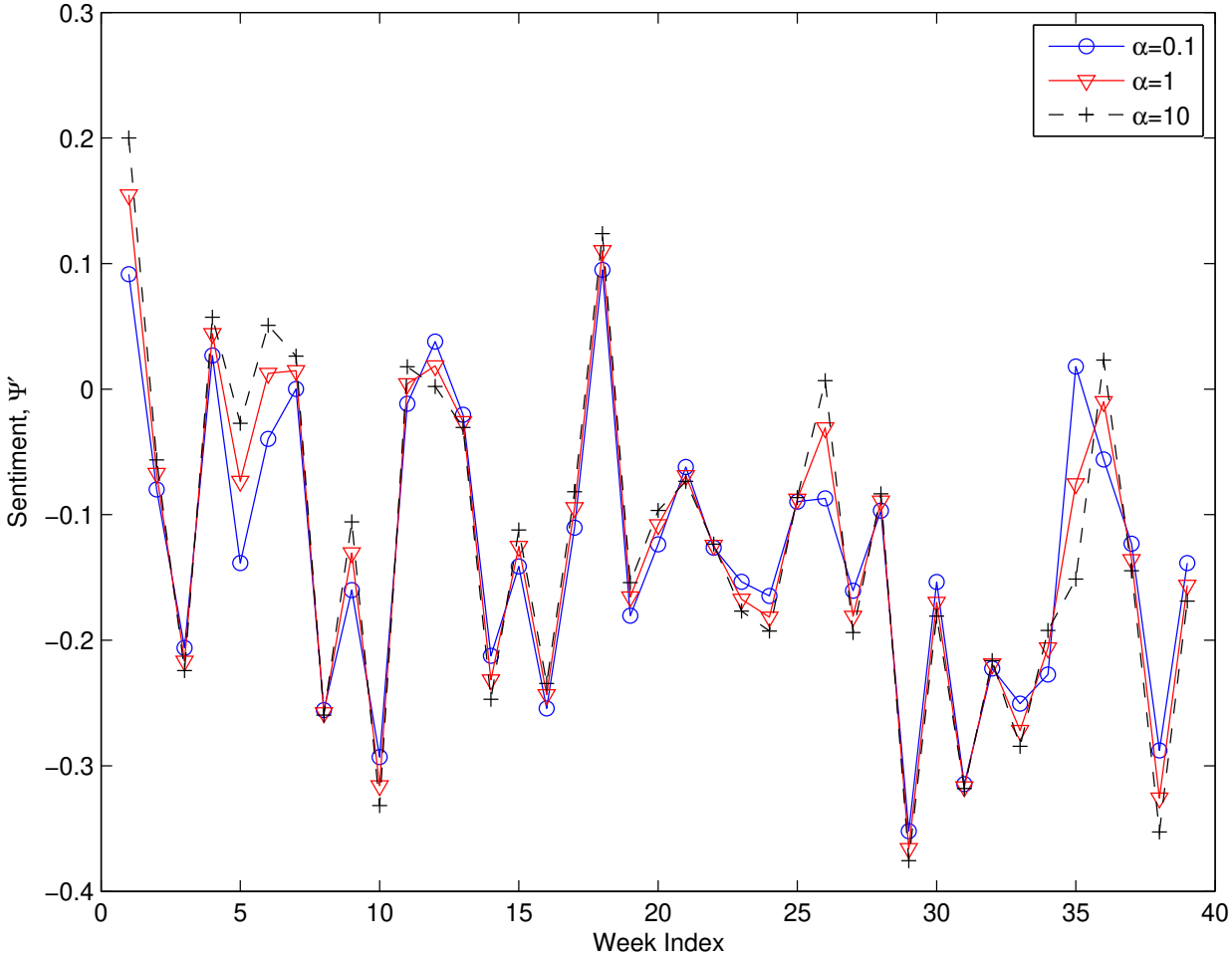

Figure III.7. Sentiment of each week, $\Psi'(k)$ with different α values.

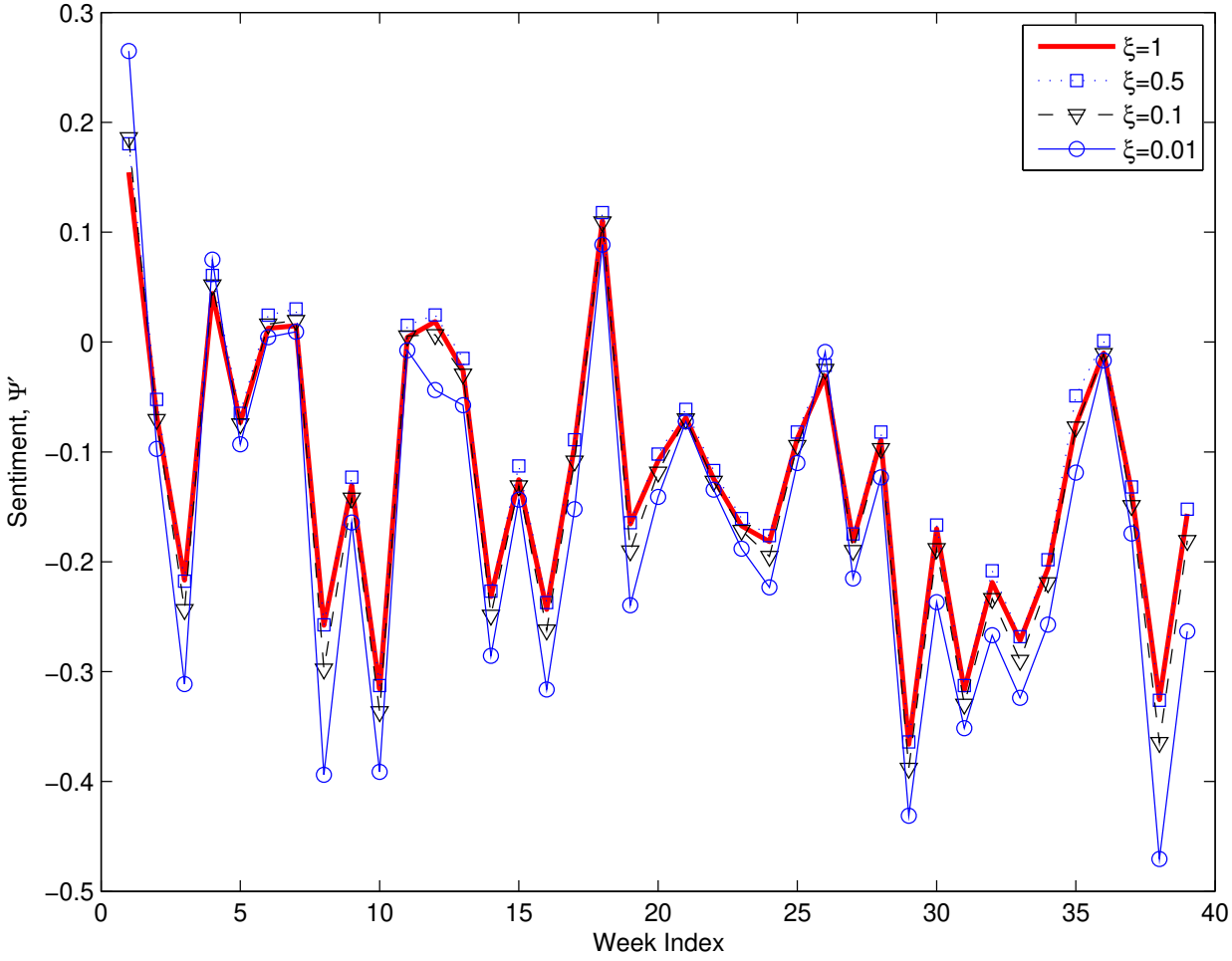

Figure III.8. Estimated sentiment based on different ξ ($\alpha = 1$).

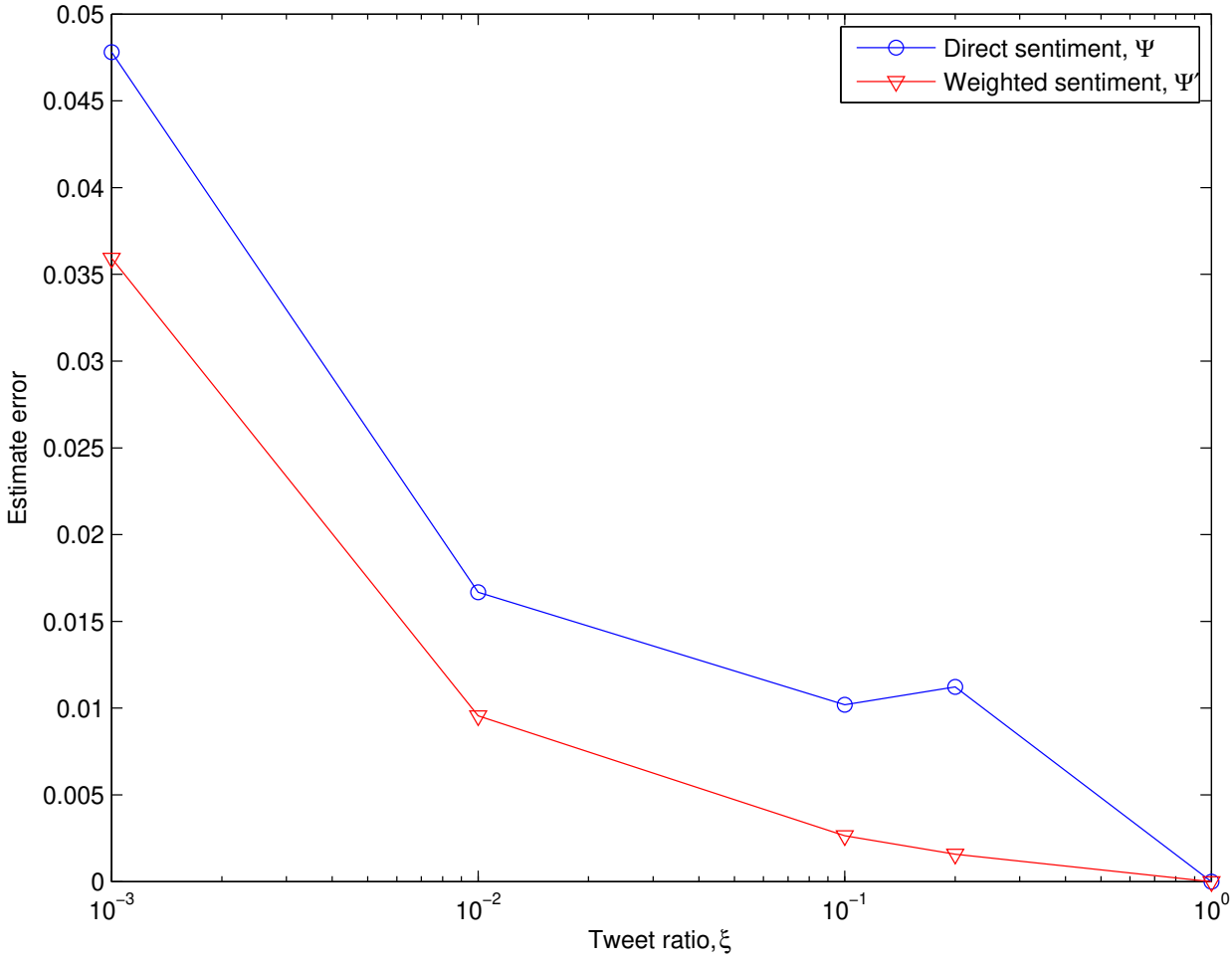

Figure III.9. Estimation error as a function of ξ ($\alpha = 1$).

CHAPTER IV

CONCLUSIONS AND FUTURE DIRECTIONS

In this work, we have crawled tweets related to one of the critical events in recent years and have analyzed them to gain insights into sentiment estimate. Firstly we have investigated the sentiments of all the tweets across 40 weeks of migrant caravan issue. We used different types of approaches and used the one with highest accuracy the Vader sentiment approach to calculate the sentiments. With these sentiments in hand we did different types of computational analysis for the scores we labeled. Even though we couldn't very properly calculate the sentiments of the emoticons we made sure it had a pretty decent accuracy. The algorithm we used seemed to have an accuracy of close to 86 percent.

The computational analysis we worked started with first analysing different categories of tweets. We have separated 40 weeks of data into individual weeks and concentrated to calculate that week in particular. We first considered the most favorite tweet of the week, was the most favourite tweet the most retweeted tweet. If yes, we have seen what is the sentiment of the this tweet, the effect of this tweets sentiment on the rest of tweets. we analyzed if a weeks sentiment had any effect on the next progressive weeks. The importance here is given to sentiment because it indeed summarizes the opinion of twitter users into a single word. Though there were many neutral sentimented tweets negative overtook over any other emotions. Also, out of 40 weeks sentiment close to 35 weeks had sentiment with negative overall sentiment and the others left to positive. We have used another type of typical calculation to see how the retweeted tweet's sentiment has effected the tweets after. The averages of before the most retweeted tweet's sentiment had more impact on the averages after the retweeted tweet's sentiment. Also, with all these computations we could conclude

that negative tweets had a greater impact than the positive tweets. Our goal here was to consider the fact that social media as an emerging body is creating a lot of strong impact on the users, we have practically investigated its impact, and could analyze the effects it had. Our results showed that a simple yet accurate estimate of the overall sentiment is to check a small portion of the most influential (defined as those with most retweets) relevant tweets. While this sounds intuitive, our results provide some numerical backings to the estimate.

It remains to be seen how the tweets with different sentiments interacted with each other, how other social media platforms affect, age groups that are getting triggered. We plan to investigate such effects in our future work. It would also be interesting to compare computation cost and accuracy of sentiment analysis with the deep learning modules in AllenNLP [22].

BIBLIOGRAPHY

[1] D. D. Luxton, J. D. June, and J. M. Fairall, "Social media and suicide: A public health perspective," *American Journal of Public Health*, vol. 102, no. S2, pp. S195–S200, 2012, pMID: 22401525. [Online]. Available: https://doi.org/10.2105/AJPH.2011.300608

[2] I. J. Emad Abu-Shanab (Information Technology College, Yarmouk University and S. A. Mushera Frehat (Al-Qassim University, Al-Qassim, "Importance of social networks the role of social networking in the social reform of young society."

[3] W. He, S. Zha, and L. Li, "A survey of data mining techniques for social media analysis," *Journal of Data Mining & Digital Humanities*, vol. 33, no. 3, pp. 464 – 472, 2013. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0268401213000030

[4] S. Elbagir and J. Yang, *Twitter Sentiment Analysis Using Natural Language Toolkit and VADER Sentiment*, Proceedings of the International MultiConference of Engineers and Computer Scientists 2019 IMECS 2019, Hong Kong, 2019. [Online]. Available: https: https://pdfs.semanticscholar.org/74a2/7879b6c245d9ff7d9c4b41175ffd84b79d73.pdf

[5] C. J. Hutto and E. Gilbert, "Vader: A parsimonious rule-based model for sentiment analysis of social media text," in *ICWSM*, 2014.

[6] L. Wang, J. Niu, and S. Yu, "Sentidiff: Combining textual information and sentiment diffusion patterns for twitter sentiment analysis," *IEEE Transactions on Knowledge and Data Engineering*, vol. 32, no. 10, pp. 2026–2039, 2020.

[7] H. Allcott, L. Braghieri, S. Eichmeyer, and M. Gentzkow, "The welfare effects of social media," *American Economic Review*, vol. 110, no. 3, pp. 629–76, March 2020. [Online]. Available: https://www.aeaweb.org/articles?id=10.1257/aer.20190658

[8] N. Al-Qaysi, N. Mohamad-Nordin, and M. Al-Emran, "A systematic review of social media acceptance from the perspective of educational and information systems theories and models," *Journal of Educational Computing Research*, vol. 57, no. 8, pp. 2085–2109, 2020. [Online]. Available: https://doi.org/10.1177/0735633118817879

[9] C. McClellan, M. M. Ali, R. Mutter, L. Kroutil, and J. Landwehr, "Using social media to monitor mental health discussions evidence from Twitter," *Journal of the American Medical Informatics Association*, vol. 24, no. 3, pp. 496–502, 10 2016.

[10] P. Burnap, O. F. Rana, N. Avis, M. Williams, W. Housley, A. Edwards, J. Morgan, and L. Sloan, "Detecting tension in online communities with computational twitter analysis," *Technological Forecasting and Social Change*, vol. 95, pp. 96–108, 2015. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0040162513000899

[11] H. Wang, D. Can, A. Kazemzadeh, F. Bar, and S. Narayanan, "A system for real-time twitter sentiment analysis of 2012 us presidential election cycle," in *Proceedings of the ACL 2012 System Demonstrations*. Association for Computational Linguistics, 2012, pp. 115–120.

[12] W. Wang, L. Chen, K. Thirunarayan, and A. P. Sheth, "Harnessing twitter" big data" for automatic emotion identification," in *Privacy, Security, Risk and Trust (PASSAT), 2012 International Conference on and 2012 International Confernece on Social Computing (SocialCom)*. IEEE, 2012, pp. 587–592.

[13] D. Davidov, O. Tsur, and A. Rappoport, "Enhanced sentiment learning using twitter hashtags and smileys," in *Proceedings of the 23rd international conference on computational linguistics: posters*. Association for Computational Linguistics, 2010, pp. 241–249.

[14] G. Gautam and D. Yadav, "Sentiment analysis of twitter data using machine learning approaches and semantic analysis," in *2014 Seventh International Conference on Contemporary Computing (IC3)*, 2014, pp. 437–442.

[15] A. Amolik, N. Jivane, M. Bhandari, and M. Venkatesan, "Twitter sentiment analysis of movie reviews using machine learning techniques," *international Journal of Engineering and Technology*, vol. 7, no. 6, pp. 1–7, 2016.

[16] O. Kolchyna, T. T. Souza, P. Treleaven, and T. Aste, "Twitter sentiment analysis: Lexicon method, machine learning method and their combination," *arXiv preprint arXiv:1507.00955*, 2015.

[17] Apoorv and Boyi, "Sentiment analyis of twitter." vol. 10, no. 2011, 2011.

[18] P. Mishra, R. Rajnish, and P. Kumar, "Sentiment analysis of twitter data: Case study on digital india," in *2016 International Conference on Information Technology (InCITe) - The Next Generation IT Summit on the Theme - Internet of Things: Connect your Worlds*, 2016, pp. 148–153.

[19] M. Anjaria and R. R. Guddeti, "A novel sentiment analysis of social networks using supervised learning," *Social Network Analysis and Mining*, vol. 4, pp. 1–15, 2014.

[20] S. Bakhshi, P. Kanuparthy, and D. A. Shamma, "Understanding online reviews: Funny, cool or useful?" in *Proceedings of the 18th ACM Conference on Computer Supported Cooperative Work & Social Computing*. ACM, 2015, pp. 1270–1276.

[21] M. Salehan and D. J. Kim, "Predicting the performance of online consumer reviews: A sentiment mining approach to big data analytics," *Decision Support Systems*, vol. 81, pp. 30 – 40, 2016. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0167923615002006

[22] M. Gardner, J. Grus, M. Neumann, O. Tafjord, P. Dasigi, N. F. Liu, M. Peters, M. Schmitz, and L. S. Zettlemoyer, "Allennlp: A deep semantic natural language processing platform," 2017.

APPENDIX A

SELECTED CODE SNIPPETS

A.1. NLTK Code

This code was more straightforward, as we used a python package that did most of the heavy lifting.

A.2. Prepossessing the Data

Data Pre Processing is the primary and the most important steps when handling raw twitter data. Text has to be cleaned before analyzing the sentiments of the tweets.

```
In [14]: import preprocessing
         import gensim
         from gensim.utils import simple_preprocess
         from gensim.parsing.preprocessing import STOPWORDS
         from nltk.stem import WordNetLemmatizer, SnowballStemmer
         from nltk.stem.porter import *
         import numpy as np
         np.random.seed(2018)
         import nltk
         nltk.download('wordnet')

         [nltk_data] Downloading package wordnet to
         [nltk_data]     C:\Users\shash\AppData\Roaming\nltk_data...
         [nltk_data]   Package wordnet is already up-to-date!

Out[14]: True

In [33]: from gensim.parsing.preprocessing import strip_multiple_whitespaces
         from gensim.parsing.preprocessing import strip_punctuation
         from gensim.parsing.preprocessing import remove_stopwords
         from gensim.parsing.preprocessing import strip_numeric

In [34]: data2.text = data2.text.str.encode('utf-8')
         df6 = data2.text.values.tolist()

In [35]: df6 = [text.lower() for text in df6]
         df6 = [strip_multiple_whitespaces(text) for text in df6]
         df6 = [strip_punctuation(text) for text in df6]
         df6 = [remove_stopwords(text) for text in df6]
         df6 = [strip_numeric(text) for text in df6]

In [36]: df8 = pd.DataFrame(df6)

In [37]: data2['text'] = df8[0]
```

Figure A.1. Pre Processing of tweets

A.3. Calculating Sentiment Polarity Scores

Calculating the sentiments of the tweets using the VADER Sentiment analysis tools.

```
In [42]: def analyser_polarity(text):
             """ Transform the output to a binary 0/1 result """
             score = analyser.polarity_scores(text)
             total_positive_score = score['pos']
             total_negative_score = score['neg']
             total_neutral_score = score['neu']
             compound_score = score['compound']

             #print "Positive Score: %s, Negative Score: %s, Neutral Score: %s, Compound Score: %s"% (pos_score, neg_score, neu_score, co

             #if score['compound'] >= 0.5:
              #   sentiment = 'Positive'
             #elif score['compound'] > -0.5 and score['compound'] < 0.5:
              #   sentiment = 'Neutral'
             #elif score['compound'] <= -0.5:
              #   sentiment = 'Negative'

             #print sentiment, ",", pos_score, ",", neg_score, ",", neu_score, ",", compound_score

             if (total_neutral_score > 1 and total_positive_score > total_negative_score and total_positive_score >= total_neutral_score)
                 sentiment = 'Positive'
             elif (total_neutral_score > 1 and total_negative_score > total_positive_score and total_negative_score >= total_neutral_scor
                 sentiment = 'Negative'
             elif (total_neutral_score > 1 and total_neutral_score > total_positive_score and total_neutral_score > total_negative_score)
                 sentiment = 'Neutral'
             elif (total_neutral_score > 1 and total_negative_score == total_positive_score and total_negative_score >= total_neutral_sco
                 sentiment = 'Neutral'
             elif (total_neutral_score <= 1 and total_positive_score == total_negative_score and total_positive_score == total_neutral_sc
                 sentiment = "Neutral"
             elif (total_neutral_score <= 1 and total_positive_score > total_negative_score):
                 sentiment = "Positive"
             elif (total_neutral_score <= 1 and total_negative_score > total_positive_score):
                 sentiment = "Negative"
             else:
                 if score['compound'] >= 0.5:
                     sentiment = 'Positive'
                 elif score['compound'] > -0.5 and score['compound'] < 0.5:
                     sentiment = 'Neutral'
                 elif score['compound'] <= -0.5:
                     sentiment = 'Negative'
             return sentiment
```

Figure A.2. calculation of sentiment polarity scores

A.4. Calculating Information Gain

code to label sentiments for sentiment scores, counting different sentiment of tweets. The overall negative tweets are way more than te positive tweets.

In [46]:
```
def f(row):
    if row['sentiment'] == 'Negative':
        val = -1
    elif row['sentiment'] == 'Positive':
        val = 1
    else:
        val = 0
    return val

df12['sentiment_values'] = df12.apply(f, axis=1)
df12.head()
```

Out[46]:

	username	date	retweets	favorites	id	text1	sentiment	sentiment_values
2	doctordiscoflor	2018-06-10 19:59	0	0	1005962758469365761	italy shuts ports migrant boat asks malta open...	Neutral	0
3	BackwardsGooner	2018-06-10 19:59	0	0	1005962748575014912	italy shuts ports migrant boat asks malta open...	Negative	-1
4	AT_Hopkins	2018-06-10 19:59	0	0	1005962690538319873	italy s matteo salvini shuts ports migrant res...	Positive	1
5	Warandhaabmedia	2018-06-10 19:58	0	0	1005962552717729792	australia took positive steps discourage migra...	Positive	1
6	DirigoPost	2018-06-10 19:58	0	0	1005962315286614017	splitting mothers kids deter migrants âœarbit...	Negative	-1

In [47]:
```
df12.sentiment.value_counts()
```

Out[47]:
```
Negative    1066283
Positive     855753
Neutral      674718
                  2
Name: sentiment, dtype: int64
```

Figure A.3. Labeling the sentiment of the tweet for the respective polarity scores

A.5. Calculate change in sentiments before and after a retweet

Retweets matching from the whole original tweets are extracted, then the average sentiments of tweets before and after the retweets are calculated.

```
In [14]: pp = {'date': [],'username': [],'avgBefore': [],'avgAfter': []}
         dff = pd.DataFrame(pp)
         a = []

         def average_low_processor(x):
             #boolean_findings = data['username'].str.contains(x['username'])
             #if boolean_findings.sum() > 0:
             grouped = data.groupby(['username']).get_group(x['username'])
             #i = np.where((grouped['date']== x['date']))
             grouped=grouped.set_index(['date'])
             #print(grouped)
             avgu =  grouped.loc[:x['date']]
             avgf =  grouped.loc[x['date']:]

             new_row = {'date':x['date'], 'username':x['username'], 'avgBefore':avgu['sentiment_values'].mean(), 'avgAfter':avgf['sentimer

             a.append(new_row)
             print ("Number of items in the list = ", len(a))



         #data1 = data1.head(10)
         data1.apply(average_low_processor,axis=1)
         data1.head()
         print(dff.size)
         dff.head()
         len(a)
         #print(a)

         fd = pd.DataFrame(a)
         fd.head(

         fd.to_csv (r'D:\SHRAVYA PROJECT\retweets_avg.csv', index=None)
```

Figure A.4. Sentiments of tweets before and after a retweet